Evaluating a user interface with ergonomic criteria

Christian Bastien, Dominique Scapin

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Evaluating a user interface with ergonomic criteria

J. M. Christian BASTIEN
Dominique L. SCAPIN

N° 2326
Août 1994

PROGRAMME 3

Evaluating a user interface with ergonomic criteria

J. M. Christian Bastien * and Dominique L. Scapin **

Programme 3
Projet Psycho-Ergo
Rapport de recherche n° 2326 - Août 1994 - 32 pages

Abstract : The usefulness of a set of ergonomic criteria for the evaluation of user interfaces was assessed using a mixed two-factors experimental design. Two groups of usability specialists (Control, Criteria) were asked to evaluate the interface of a musical database management system in two phases. In the first phase of the experiment, all the participants relied solely on their expertise; in the second phase they were instructed to evaluate the management system again but this time through the replay of their previous interactions: participants in the Criteria group used a set of ergonomic criteria while the participants in the Control group did not. In the first phase, the two groups did not differ in terms of: (1) the number of usability problems detected, and (2) the proportions of usability problems uncovered as well as the proportion of usability problems found in common, with respect to the size of the aggregates. In the second phase however, the participants in the Criteria group had better performances than those in the Control group: they uncovered more new problems, and the proportion of problems uncovered as well as the proportion of problems found in common was greater as a function of the number of evaluators in the aggregates. To sum up, the criteria increased the evaluation performance of the experts.

Key-words : User interface evaluation, heuristic evaluation, ergonomic criteria, standards, usability problems, usability expertise, cost-effective methods.

This research was supported in part by a grant from the European Software Factory. Part of this study was presented in a poster format at the 1993 Conference on Human Factors in Computing Systems (INTERCHI'93), Amsterdam, The Netherlands.

* E-mail : Christian.Bastien@inria.fr
** E-mail : Dominique.Scapin@inria.fr

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)
Téléphone : +33 (1) 39 63 55 11 - Télécopie : +33 (1) 39 63 58 88 - Télex : 697 033 F

Évaluer une interface utilisateur avec des critères ergonomiques

Résumé : L'utilité d'un jeu de critères ergonomiques pour l'évaluation des interfaces utilisateur a été appréciée à l'aide d'un plan expérimental mixte à deux facteurs. Deux groupes de spécialistes en ergonomie des logiciels (Contrôle, Critères) ont été invités, au cours de deux phases, à évaluer l'interface d'un système de gestion d'une base de données musicales. Dans la première phase de l'étude, tous les participants ont évalué l'interface en ne se basant que sur leur propre expertise. Au cours de la deuxième phase de l'étude, les participants ont évalué l'interface une deuxième fois mais cette fois-ci en visionnant l'enregistrement de leurs interactions précédentes. Les participants du groupe Critère, contrairement à ceux du groupe Contrôle ont utilisé pour ce faire le jeu de critères ergonomiques. Au cours de la première phase, aucune différence n'a été observée entre les groupes quant (1) au nombre de problèmes d'utilisabilité détectés et, (2) quant à la proportion de problèmes détectés et la proportion de problèmes détectés en commun selon le nombre d'évaluations réunies. Au cours de la deuxième phase toutefois, les performances des participants du group Critères se sont avérées supérieures à celles des participants du groupe Contrôle. Les premiers ont détectés plus de nouveaux problèmes que les seconds et la proportion des problèmes détectés de même que la proportion des problèmes détectés en commun est apparue supérieure en fonction du nombre d'évaluation mis en commun. En résumé, les critères ont amélioré les performances d'évaluation des experts.

Mots-clé : Évaluation d'interfaces utilisateur, évaluation heuristique, Critères ergonomiques, standards, problèmes d'utilisabilité, expertise en utilisabilité.

1 Introduction

Expert-based or heuristic evaluation is generally defined as an informal method of usability analysis where experts (human factors specialists, system designers, software engineers, etc.) are presented with an interface and asked to comment on it (e.g., Nielsen & Molich, 1990), i.e. to identify the ergonomic design flaws. In this type of evaluation, the experts rely on their experience and heuristics in order to make a judgement on the ergonomic quality of the system, or they appraise the conformance of the system with established human factors standards, guidelines, principles, and criteria.

Even though there can be no single best evaluation method, expert evaluation is relatively fast, and uses few resources. It has been found to provide a more integrated view than an evaluation based on user performance (Hammond, Hinton, Barnard, MacLean, Long, & Whitefield, 1985), and to be superior to the application of guidelines or to cognitive walkthroughs both in terms of amount and importance of problems found, and cost/benefit ratio, i.e. the number of problems found per person-hour (Jeffries, Miller, Wharton, & Uyeda, 1991). Furthermore, a few experts have been found to uncover more problems than any other method and to predict about half of the problems found in a usability test (Desurvire, Kondziela, & Atwood, 1992).

These advantages however, should not be interpreted as a proof that heuristic evaluation is more efficient and effective than usability testing and that the latter is a "waste of time" (Bevan, 1991; Tognazzini, 1992 cited in Jeffries & Desurvire, 1992). On the contrary, "a single heuristic evaluation was consistently the least powerful evaluation technique" (Jeffries & Desurvire, 1992, p. 39). It missed about half of the problems found with usability testing, and this latter technique missed about the same number of problems found in heuristic evaluation (Desurvire, et al., 1992). Heuristic evaluation was better only when multiple evaluations were aggregated. Finally, usability testing uncovered more severe problems, more recurring problems and more global problems than did the heuristic evaluation (Jeffries, et al., 1991). In addition, studies that compared heuristic evaluation with other methods have been shown to be in internal conflict and in disagreement with one another when compared using criteria such as the average cost in finding problems, and the proportion of severe problems found by each method (Muller, Dayton, & Root, 1993). It seems that any conclusions regarding overall superiority of one method with respect to another are premature (Muller, et al., 1993).

Although the comparison of heuristic evaluation with other methods is necessary to determine the advantages and drawbacks of each method, heuristic evaluation per se must be investigated if one wants to provide the evaluators with more usable tools and materials, and with methods that will guaranty more precise and more reliable data. Ideally, those tools, materials and methods should permit to reduce the variability of the evaluators' performances in relation to their academic background and the type of target system evaluated.

The materials currently available to the experts are basically standards (e.g., AFNOR, ANSI, DIN, ISO, etc.), general design guides (e.g., Brown, 1988; Ravden & Johnson, 1989; Scapin, 1986; Shneiderman, 1987, etc.), sets of guidelines (e.g., Bodart & Vanderdonckt, in press; Smith & Mosier, 1986), algorithms (e.g., de Baar, Foley, & Mullet, 1992; Vanderdonckt & Bodart, 1993), style guides (e.g., Apple Computer Inc.,

1989; IBM, 1989), checklists (e.g., Oppermann, Murchner, Paetau, Pieper, Simm, & Stellmacher, 1989; Ravden & Johnson, 1989), and heuristics/criteria (e.g., Bastien & Scapin, 1993; Molich & Nielsen, 1990; Nielsen, 1994; Scapin, 1990a, 1990b).

All of these documents have been developed for the purpose of good human-computer interface design. Paradoxically, only a few of these documents have been evaluated in terms of their reliability, their validity, their thoroughness, their effectiveness, and their ease of use by usability specialists as well as non-specialists. Apparently none has been evaluated regarding its completeness in relation to the available knowledge in the field of human-computer interaction.

Guidelines have been found to be difficult to use. These difficulties, observed in design tasks as well as in evaluation tasks, were related to locating the relevant guidelines and establishing priorities among them (Mosier & Smith, 1986), to the interpretation of the guidelines, the identification of the design problems and to their solutions (de Souza & Bevan, 1990; Tetzlaff & Schwartz, 1991; Thovtrup & Nielsen, 1991).

An alternative to the use of large sets of guidelines has been the use of heuristics and/or criteria. The sets of heuristics/criteria currently available vary from one author to another. The variability in the degree of specificity, in the number and precision of heuristics/criteria depends on the way they were defined. Different design strategies seem to have formed the basis for the currently published heuristics/criteria. One of these has been based on an examination of knowledge derived from research on high-level cognitive processes such as reasoning, mental models, memory, language, and skill acquisition. From such knowledge, recommendations were extracted and organised into high level criteria or dimensions (e.g., Marshall, Nelson, & Gardiner, 1987). Another design strategy has been based on the review of currently published criteria with the goal of organising them into a common structure (e.g., Johnson, Clegg, & Ravden, 1989; Ravden, 1988; Ravden & Johnson, 1989). Sometimes personal experience has been coupled with existing principles (e.g., Molich & Nielsen, 1990), and several published sets of usability heuristics have been compared with a database of usability problems in order to determine which heuristics best explained them (e.g., Nielsen, 1994). Another design strategy has been more empirical: it started from available experimental data and recommendations that were translated into rules, then iteratively grouped into sets that were characterised by specific criteria (Scapin, 1990a, 1990b).

The design of valid and reliable heuristics/criteria is essential for: the transfer of human factors knowledge to designers; the structuring of training in human factors; the design of evaluation grids and the organisation of evaluation reports; the design of metrics; the retrieval of recommendations in human factors data bases, and the design of computer-based evaluations. However, in order to achieve these goals, two main questions must be raised: (1) how usable are these different sets, and (2) in what ways do the different sets of heuristics/criteria relate to the available recommendations. In other words, the sets of heuristics/criteria must be defined explicitly, unambiguously, consistently, and they must be closely related to individual recommendations.

The set of heuristics proposed by Molich and Nielsen (1990) has been used, in several experiments, by usability specialists as well as non-specialists to evaluate different systems. It has been shown that individual evaluators were "quite bad" at evaluating different systems with the heuristics. They only found between 20 and 51%

of the usability problems the interfaces contained (Nielsen & Molich, 1990). On the other hand, when the reports of several evaluators were aggregated into a single evaluation, such aggregates did rather well even with only three to five people (Nielsen & Molich, 1990). Furthermore, Nielsen (1992) noted that usability specialists were better than those without usability expertise at finding usability problems by heuristic evaluation. Usability specialists with expertise in the specific kind of interface being evaluated were also much better than regular usability specialists without such expertise, especially with regard to certain usability problems that were unique to the specific kind of interface. It also appeared that major usability problems had a higher probability of being detected than minor problems in a heuristic evaluation, but that about twice as many minor problems were found in absolute numbers, and some problems (e.g., lack of clearly marked exits) appeared to be more difficult to uncover than others.

The experiments from which the previous results were taken did not explicitly state the procedures the participants were instructed to follow when applying the heuristics: were the participants instructed to read the heuristics before the evaluation and then given free access to the heuristics during their task? Were they instructed to evaluate the systems using the heuristics one by one? How were the heuristics actually used? Were they used as a classification tool for problems the participants uncovered by means of their experience, or were the heuristics really responsible for the detection of usability problems? Did the usability problems relate directly to the heuristic being used or were the heuristics just occasions upon which usability problems in general were uncovered?

All the previous questions concern the use of heuristics/criteria as an evaluation method and to the characteristics such a method should possess. An evaluation method based on heuristics/criteria should be, at least: valid, i.e. permitting the expert to evaluate systems on those aspects the heuristics/criteria are intended to evaluate; thorough, i.e. allowing the widest scope of the interface evaluation as possible; and reliable, i.e. providing the same results under the same conditions.

To develop an evaluation method based on ergonomic criteria that have the aforementioned characteristics and, to ensure the validity of the set as well as its thoroughness, Scapin (1990a, 1990b) adopted an empirical approach and proposed a set of ergonomic criteria. This set was first used informally by Pollier (1992) who demonstrated that it allowed an accurate description and classification of the usability problems found by experts in an evaluation task.

The reliability of this set was investigated experimentally in an identification task in which human factors specialists and non-specialists were asked to identify the criterion, within the set, that was violated for each of the usability problems they were provided with (Bastien & Scapin, 1992). In this study, statements of usability problems that were presented to the participants included explanations about the interaction context, and were illustrated by copies of display screens. Usability problems were thus already identified: participants had only to match them with the corresponding criteria. No difference was observed between groups of participants either in terms of performance times or correct matches. A detailed examination of the data and an analysis of confusion matrices lead to improvements of the set of criteria: some criteria were modified and, examples, counter-examples and comments were added.

Another step in the development of the evaluation method was to assess the usefulness, for usability specialists, of this set of ergonomic criteria in an evaluation task. In other words, we were interested in assessing what the set of ergonomic criteria would add to an evaluation relying solely on expertise.

In order to give usability specialists interaction opportunities, a fully implemented system was used instead of a paper implementation of the interface consisting of a series of screen designs.

2 Method

2.1 Participants

Twenty usability specialists were randomly assigned either to the "Control" or to the "Criteria" group. The data from one participant in the Criteria group were lost due to equipment failure. There were thus 10 participants in the Control group, and 9 in the Criteria group. All the usability specialists had at least a post-graduate degree in ergonomics[1] except one in the Control group who had a master's of art.

On average, participants had 6 (SD = 3.8), and 4.56 years of experience (SD = 2.19) in the evaluation and/or design of user interfaces in the Control and Criteria group respectively (F (1, 17) = 1 ; p = .3316) (See Table A1 and A2 in the Appendix section for more details on the ANOVA and Means tables).

The data gathered from a questionnaire indicated that the participants in the two groups were quite similar with respect to their academic background, the type of evaluation conducted, and their computer experience. To summarise, most of the participants reported devoting about 30% of their time to the evaluation of user interfaces. They all had experience in the evaluation of specifications, screen dumps, mock-ups, prototypes and operational systems. When conducting their evaluation the specialists reported that they relied greatly on their own expertise. They all resorted to user testing to some extent, and they made only little use of guides and/or recommendations.

All the participants were familiar with the Apple Macintosh, and (except three participants in both groups) were either HyperCard™ stack users and/or programmers (See Table A2 in the Appendix section for more details).

2.2 Materials and Equipment

The set of ergonomic criteria. The set of ergonomic criteria[2] the participants in the Criteria group were asked to use in the second phase of their evaluation was composed of eighteen elementary criteria, some of which are grouped under more general ones (see Table 1). Each elementary criterion was presented in a document along with a definition, a rationale, examples of recommendations, and comments.

---

1 The post-graduate degrees consist of a DESS and/or a DEA which correspond to the fifth year in the French university programs.

2 The set of ergonomic criteria (with their definitions, rationales, examples, and comments) may be obtained from the authors.

Table 1. ***List of the Ergonomic Criteria*** (*)

1 Guidance
- 1.1 **Prompting (PROM)**
- 1.2 Grouping and distinction of items
  - 1.2.1 **Grouping and distinction of items by location (GDLO)**
  - 1.2.2 **Grouping and distinction of items by format (GDFO)**
- 1.3 **Immediate feed-back (FEED)**
- 1.4 **Legibility (LEGI)**

2 User workload
- 2.1 Brevity
  - 2.1.1 **Concision (CONC)**
  - 2.1.2 **Minimal actions (MIAC)**
- 2.2 **Information density (INDE)**

3 User explicit control
- 3.1 **Explicit user actions (EXUA)**
- 3.2 **User control (USCO)**

4 Adaptability
- 4.1 **Flexibility (FLEX)**
- 4.2 **Users' experience management (USEX)**

5 Error management
- 5.1 **Error protection (ERPR)**
- 5.2 **Quality of error messages (QUEM)**
- 5.3 **Error correction (ERCO)**

6 **Consistency (CONS)**

7 **Significance of codes (SICO)**

8 **Compatibility (COMP)**

(*) Elementary criteria are in boldfaced type and are followed by their abbreviations

The management system. The system to be evaluated was a musical database management system, "Hypermanip" (Scapin, 1991) developed with HyperCard™ for the purpose of the experiment. This screen-based graphical user interface was designed so as to allow users to manage their musical library, i.e., to add new musical records, to retrieve and modify them. This area of application was chosen because the functionalities were believed to be familiar to most of the participants, to be not too complex and therefore to require no training. Figure 1 illustrates some of the display screens of the application.

Hypermanip was purposely designed to contain usability problems. While designing the system, attention was given to ensure that each ergonomic criterion be represented by a certain number of design problems in different display screens. A first list of these usability problems was drawn up. Then the interface was re-examined by the experimenters and usability problems that went unnoticed during the design process were added to the list. Though the system was well known to the experimenters and though each state of the system had been carefully reviewed, the list of usability problems was still incomplete. It had to be refined in light of the participants' evaluations. The final list contains a total of 503 instances of usability problems. The number of usability problems per criterion ranges from a minimum of 5 (e.g., "User experience") to a maximum of 99 (e.g., "Consistency"). This rather large number of Consistency problems represents each and every instance of all the possible comparisons made for any given consistency problem. Because consistency problems imply several comparisons, the number of instances grows rapidly. For example, two labels, "Category" and "Domain" were used for the data field corresponding to the music category (e.g., opera, movie songs, etc.) in three different display screens.

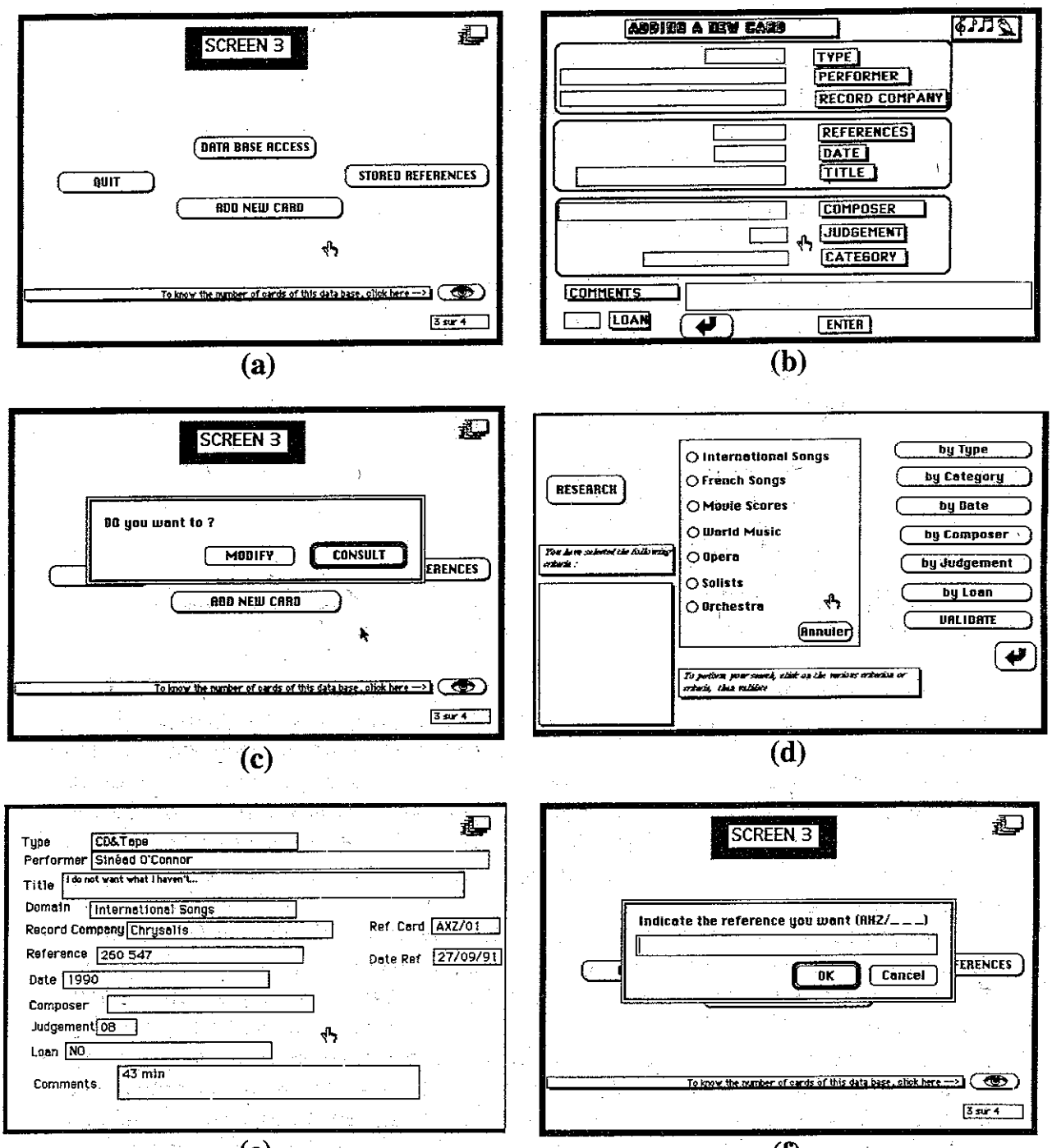

Figure 1. **Examples of display screens found in Hypermanip: (a) the main screen, (b) the screen for the addition of new records, (c) the dialogue box appearing when the user wants to access the database, (d) the screen allowing the selection of search options, (e) the records as they are presented to the participants in a computer-paced manner, (f) the dialogue box allowing the participant to get a record for modification.**

From an evaluation point of view, this problem of consistency could be pointed out while comparing display A and B, A and C, B and C, and finally A, B and C. In this case there were thus four possible instances.

The apparatus. Hypermanip was run on a portable Macintosh. All the participants' inputs as well as all the interactions were recorded on line with MediaTracks™, which allowed the replay of all interactions.

A video camera was used to capture the computer screen and to record the comments of the participants. The videotapes were later used to collect the problems uncovered by the participants.

2.3 Design

A mixed two-factor design was used in this experiment with "Group (2)" as the between-subjects factor and "Phase (2)" as the within-subjects factor. The first phase of the experiment was identical for both groups (Control and Criteria), and consisted of evaluating the interface. In the second phase of the experiment, the participants of both groups were asked to carry on their evaluation through the replay of their previous interactions with the system. In this second phase, the participants in the Criteria group used the set of ergonomic criteria while the Control group did not.

2.4 Task and Procedure

The participants were tested individually. The instructions they were given prior to the experimental session told them briefly what the management system was intended to do and informed them that it already contained a few records. In order to improve the system, they were asked to evaluate it as completely and as precisely as possible. They were also instructed to mention to the experimenter each and every ergonomic flaw they could diagnose, even when these design flaws could be considered minor. The experimenter only intervened to encourage greater precision when the comments of the evaluators were not clear or precise enough.

Phase 1. Following the reading of the instruction sheet, the participants were asked to begin their evaluation. Their walkthrough, i.e. their inputs as well as their mouse cursor movements were recorded with MediaTracks™, and their comments along with the computer screen were recorded on video. Their evaluation was self-guided, and limited to one hour.

At the end of this phase, and before the second phase, the participants in the Criteria group were given the set of ergonomic criteria and asked to read it very carefully.

Phase 2. In the second phase of the experiment, the participants of both groups were asked to evaluate the interface again but this time through the replay of their previous walkthrough. The replay was used to ensure that the diagnoses of the participants would concern exactly the same screens and dialogue boxes that had been explored in the first phase. New problems uncovered in the second phase could therefore be attributed to a second look at the same screens and dialogue boxes rather than to the exploration of new ones. The replay was paused at each recorded interaction state. The pauses lasted the time the participants needed to make their comments, then the replay proceeded forward.

All the participants were told not to ask themselves if they had uncovered the problems in the first phase but once again to mention each and every problem as they uncovered them. The duration of the second phase was determined by the time taken by the participants to comment on the replay of their previous walkthrough.

In the Control group, the participants proceeded exactly as in the first phase. In the Criteria group, the participants were asked to use and apply each of the 18 elementary criteria systematically one by one for each state of the system. This instruction aimed at ensuring that the criteria would be used to diagnose the design problems rather than to classify them. During this second evaluation, the experimenter prompted these participants to refer to the document containing the criteria when they tended not to use them.

At the end of the session, the participants were presented with a questionnaire about their academic background, experience with computers, etc.

2.5 Data collection and analysis

The primary data of this experiment were the participants' verbal descriptions of the usability problems recorded on videotape during their walkthrough. Each tape was reviewed and coded with the help of the list of usability problems.

3 Results

In the first phase of the experiment some participants took advantage of the entire hour and explored all the functionalities of the system: 33.3% in the Criteria group, and 10% of the participants in the Control group.

Other participants also took advantage of the entire hour but did not explore the modification functionality: 33.3% of the participants in the Criteria group, and 70% of the participants in the Control group.

Few participants stopped before the end of the first phase telling the experimenter they had completed their evaluation, and explored all the functionalities: 2 participants in the Criteria group (22.2%), and none of the participants in the Control group.

Finally, 1 participant in the Criteria group (11.1%), and 2 in the Control group (20%), stopped before the end and did not explore the modification functionality.

The time taken by the participants who stopped before the end was 39, 42, and 55 min respectively (M = 44.3 min) for the 3 participants in the Criteria group, and 42 and 44 min respectively (M = 43 min) for the 2 participants in the Control group.

3.1 Number of problems uncovered

Figure 2 shows the cumulated number of problems uncovered for both groups across phases. A multivariate analysis of variance for repeated measures was computed with group (2) and phase (2) as main effects. The analysis indicates a significant effect for group ($F(1, 17) = 5.59$; $p = .0302$), phase ($F(1, 17) = 333.85$; $p = .0001$), and their interaction ($F(1, 17) = 10.36$; $p = .005$) (See Table A4 through A7 in the Appendix section for more details on the MANOVA and Means tables).

Globally, the participants in the Criteria group uncovered more problems (M = 89.94 ; SD = 26.20) than the participants in the Control group (M = 77.75 ;

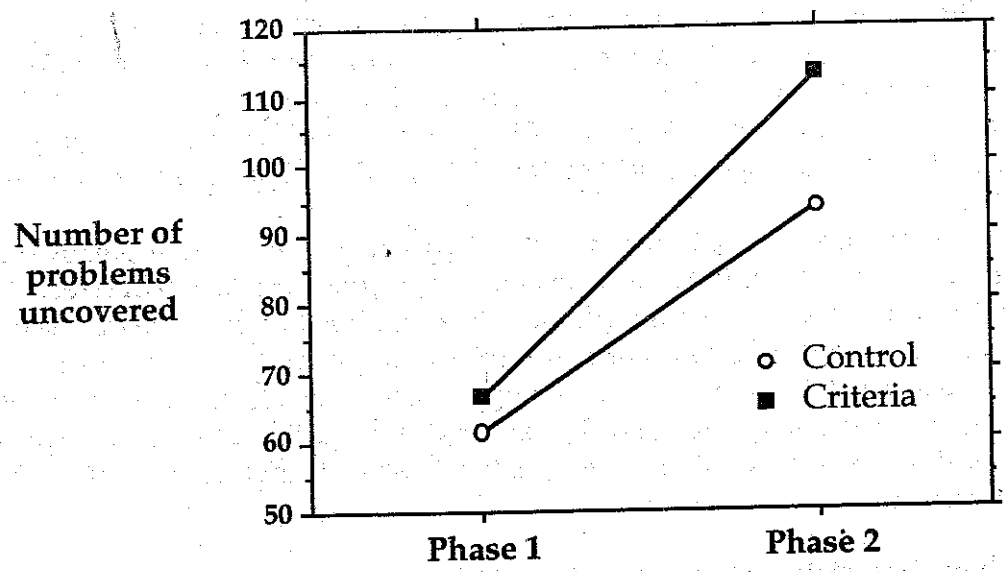

Figure 2. **Mean cumulated number of problems uncovered in each group across phases.**

SD = 20.70). The number of problems uncovered in phase 1 (*M* = 64.11 ; *SD* = 10.87) increased significantly in the second phase (*M* = 102.95 ; *SD* = 16.26). The significant interaction of group and phase effects indicates that the increase in the number of cumulated problems from phase 1 to phase 2 is significantly different in the Criteria group (see Figure 2). On average, participants in this latter group uncovered more new problems (*M* = 46.11 ; *SD* = 11.47) during the second phase than did participants in the control group (*M* = 32.3 ; *SD* = 6.91).

Interaction contrasts computed within univariate analyses of variance indicate that in phase 1 the number of problems uncovered by evaluators in the Control group (*M* = 61.6 ; *SD* = 10.48) did not differ significantly from the number of problems uncovered by evaluators in the Criteria group (*M* = 66.89 ; *SD* = 11.21) (*F* (1, 17) = 3.038 ; *p* = .0994). On the other hand, the difference is significant in the second phase (*F* (1, 17) = 39.618 ; *p* = .0001) where the mean cumulated number of problems uncovered by the participants is 93.9 (*SD* = 14.66) and 113 (*SD* = 11.69) in the Control and Criteria group respectively (See Table A8 through A12 in the Appendix section for more details on the interaction contrast tables).

To ascertain that the previous results were not related to the number of participants who did not walk through the modification functionality and/or finished their evaluation before the end of the first phase, two multivariate analyses of variance for repeated measures were computed: one with completion of the first phase (2), group (2), and phase (2) as main effects, and the other one with exploration of the modification functionality (2), group (2), and phase (2) as main effects (See Table A13 through A22 in the Appendix section for more details on the MANOVA and Means tables). The first analysis indicates a marginal effect for group ($F(1, 15) = 4.51$; $p = .0507$), a significant effect for phase ($F(1, 15) = 278.75$; $p = .0001$), and a significant interaction of these two factors ($F(1, 15) = 11.15$; $p = .0045$). Completion of the first phase had no significant effect ($F(1, 15) = 2.73$; $p = .1190$) on the mean number of cumulated problems uncovered. All the other interactions were not significant.

The second analysis (exploration of the modification functionality X group X phase) also indicates significant effects for group ($F(1, 15) = 8.04$; $p = .0125$), phase ($F(1, 15) = 163.7$; $p = .0001$) and their interaction ($F(1, 15) = 6.21$; $p = .0249$), but no significant effect ($F(1, 15) = 3.45$; $p = .0829$) of the modification functionality on the number of cumulated problems uncovered. All the other interactions were not significant.

To summarise, the performance of the participants from both groups did not differ, in the first phase, in terms of the number of problems they uncovered. In the second phase however, even though the cumulated number of problems uncovered increased significantly for the participants from both groups, this increase was greater for the participants in the Criteria group. Further analyses revealed that these conclusions could not be invalidated by the fact that the participants did not explore the modification functionality of the system and/or finished their evaluation before the end of the first phase.

3.2 Number of problems uncovered per elementary criteria

Figure 3 shows the number of problems uncovered, by the participants, per elementary criteria. This figure shows that many design flaws related to some criteria went unnoticed. In phase 1 as well as in phase 2, very few of the problems related to error management were uncovered (Error protection (ERPR), Quality of error messages (QUEM), Error correction (ERCO)). The case is similar for design flaws related to User experience management (USEX), Flexibility (FLEX), Concision (CONC), and Explicit user action (EXUA).

Among the design flaws that were uncovered in phase 1 (Figure 3), two sets may be roughly delineated in terms of the number of problems uncovered. The first group of criteria, those for which more design flaws were uncovered, is made up of the criteria Concision (CONS), Compatibility (COMP), Prompting (PROM), Minimal action (MIAC), and Significance of codes (SICO). The second group of criteria, those for which few design flaws were uncovered, comprises the criteria Information density (INDE), Grouping and distinction of items by location (GDLO) and format (GDFO), User control (USCO), Feedback (FEED), and Legibility (LEGI). The first group of criteria is the one that benefited most from the second phase of the experiment: the number of problems associated to each of these criteria increased noticeably in comparison to the second group of criteria.

But most importantly, as shown in figure 3, the distribution of the number of problems uncovered per elementary criteria in both phases is quite similar for both groups, even though participants in the Criteria group generally uncovered more problems per elementary criteria in the second phase. The effect of the set of ergonomic criteria seems to have been global. In addition, in phase 2, the distribution of design flaws per elementary criteria tends to follow the distribution of design flaws contained in the system let alone those related to Error protection.

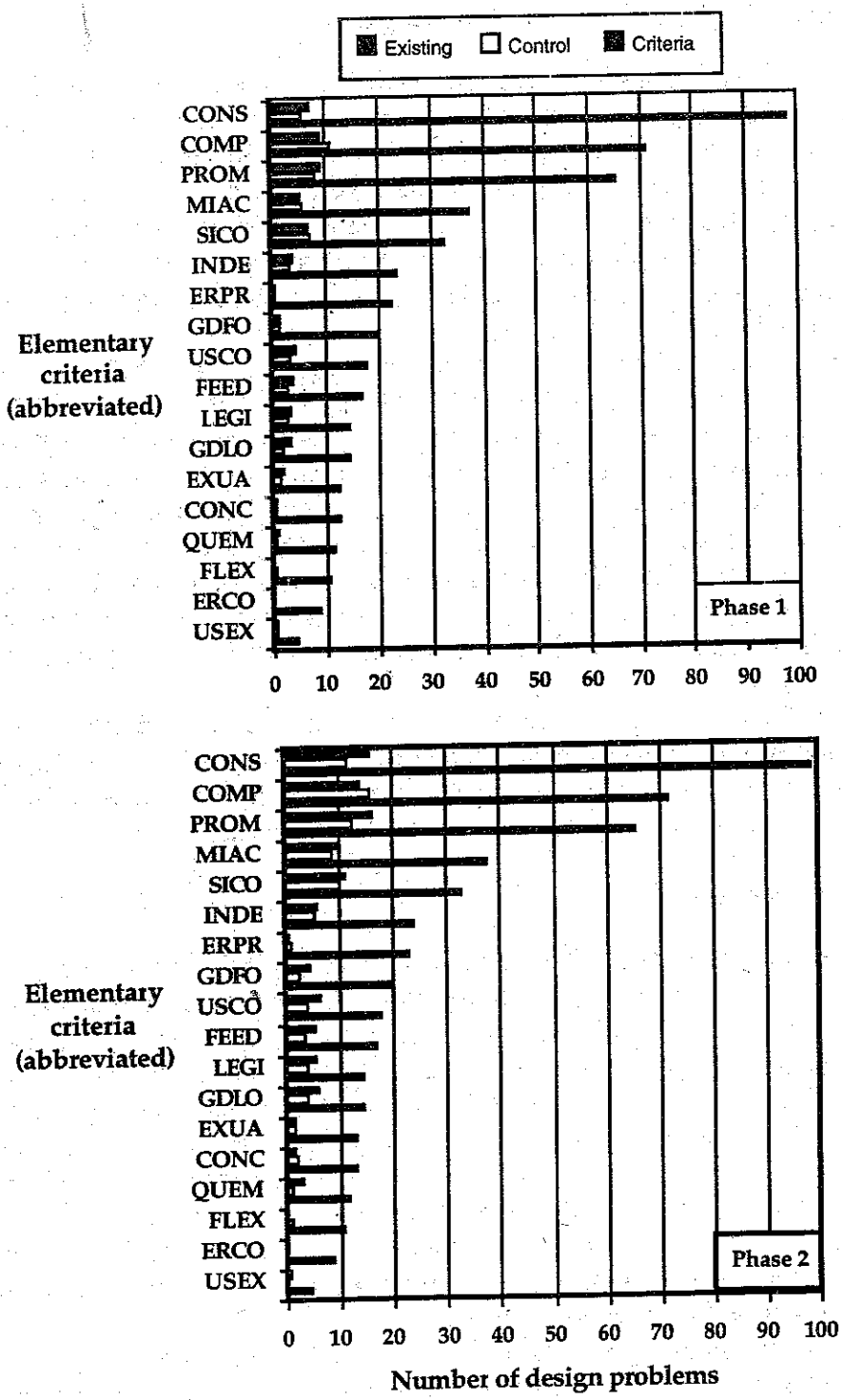

Figure 3. **Mean cumulated number of problems uncovered per elementary criteria in each group for each phase, and total number of problems existing in the management system.**

3.3 The influence of the criteria on the diagnoses

The distribution of problems uncovered per elementary criteria was obtained by a classification made by the experimenters. When the participants made their evaluation, the problems they reported were not necessarily formulated in terms of the set of ergonomic criteria. Only in the second phase of the experiment did the participants in the Criteria group report the problems in those terms. As it turned out, some criteria led these participants, on some occasions, to detect problems the experimenters had classified under some other elementary criteria. Moreover, some design flaws were identified without explicit reference to the criteria.

In the second phase of the experiment, participants in the Criteria group uncovered a mean of 83.78 problems (SD = 11.87). On average, the detection of 12.11 (SD = 5.67) out of these problems (14%) was induced by elementary criteria that were not supposed to do so. For example, when evaluating the legibility of a display screen on which there was a feedback message, some evaluators stated that the inappropriate size and style of the fonts was a legibility problem. According to the definition given for the Legibility and Immediate Feedback criteria, this design flaw should not have been induced by the Legibility criterion but by the Immediate Feedback criterion. These are the kinds of "misguidance" that are important to note in order to refine the definitions of the criteria.

Some design flaws were identified without explicit reference to the criteria. These are the cases where the participants tended to put aside the list of ergonomic criteria during their evaluation. An average of 6.22 problems (SD = 2.28) out of 83.78 (8%) were identified without any explicit reference to the elementary criteria.

3.4 Aggregated evaluations

Aggregates of evaluators were formed by collecting the design flaws uncovered by different number of evaluators to form larger sets. The number of evaluators per aggregate varied from 1 to the total number of evaluators in each group. For each number of evaluators in the aggregate, the number of selections of "r" participants taken out of the "n" participants with no attention given to the order of arrangement is given by $\frac{n!}{r!(n-r)!}$. For example, there are 126 possible combinations of participants taken 4 by 4 out of nine in the Criteria group, and 210 combinations of participants taken 4 by 4 out of ten in the Control group. For each combination the number of problems uncovered was determined. This number was then divided by the total number of known problems, i.e. 503, to give the proportion of problems uncovered. Means were then calculated on these proportions.

The mean proportions of ergonomic problems found for each size of aggregates of evaluators are shown in Figure 4. This figure shows that in phase 1 both groups present an identical trend. On average participants uncovered 12% and 13% of the total number of design flaws of the management system in the Control and the Criteria group respectively. By aggregating the evaluation of all the participants within a group, 48% of the problems were uncovered by the Control group for 46% for the Criteria group.

In the second phase of the experiment, the proportions of problems uncovered as a function of the size of the aggregate are different for the two groups. Individually, participants in the Control group uncovered an average of 19% of the design flaws while the participants in the Criteria group uncovered a mean of 22%.

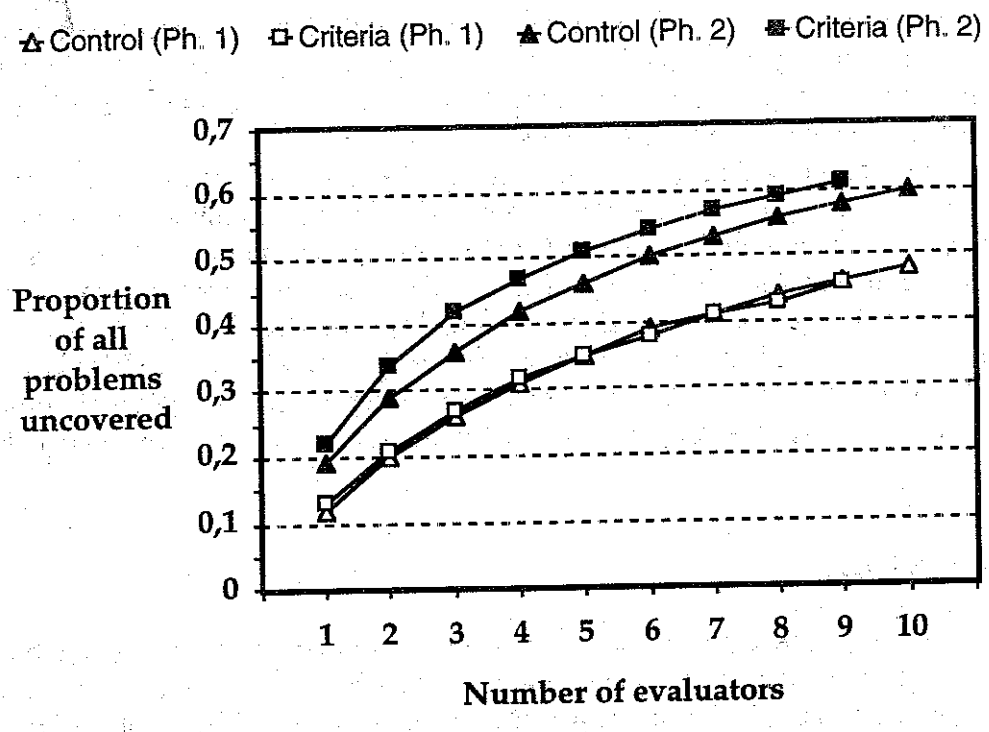

Figure 4. **The mean proportion of ergonomic problems found in each phase as a function of the number of evaluators in the aggregates.**

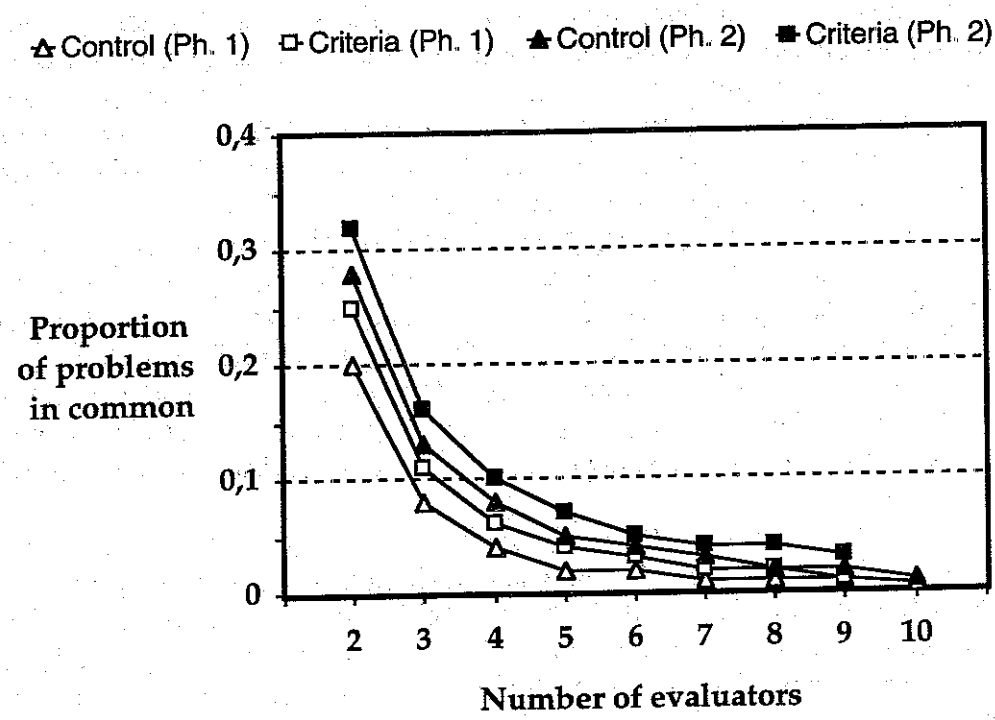

Figure 5. **The mean proportion of ergonomic problems found in common in each phase as a function of the number of evaluators in the aggregates.**

The difference is not very large but when one looks at the evolution of the curve one finds that uncovering about 50% of the known design flaws required 5 participants in the Criteria group against 6 or 7 in the Control group. In other words, the set of ergonomic criteria would tend to necessitate less evaluators to get the same results as compared to a evaluation conducted without it.

Another effect of using or not using the set of ergonomic criteria is reflected in the proportion of design flaws the participants uncovered in common. As shown in Figure 5 the tendency is quite the same for the two groups: the proportion of design flaws detected in common is relatively small and decreases very rapidly as the number of evaluators in the aggregate increases. Although these proportions are higher in the second phase of the experiment for both groups, the participants in the Criteria group produce the highest proportions. In other words, using the set of ergonomic criteria would tend to increase the similarity of the evaluations more than not using them would.

4 Discussion

The aim of the present paper was to assess the usefulness of the set of ergonomic criteria for the evaluation of a human-computer interface. The results indicate that usability specialists benefited from the use of ergonomic criteria. The number of design flaws uncovered in the second evaluation of the interface was greater for the evaluators who used the criteria than for those who relied solely on their expertise. This was reflected in aggregates of evaluations: the number of evaluators necessary to uncover a specific proportion of design flaws was less for the criteria group than for the control group. In addition, the use of ergonomic criteria produced evaluation reports that were more similar: this was reflected in the proportions of problems found in common as a function of the number of evaluators in the aggregate. Some usability problems appeared to be more difficult to uncover than others, even with the help of ergonomic criteria. Overall, the participants made very few errors of classification.

Some caution must be taken however when interpreting the results. In the first phase of the experiment, which was limited to one hour, the evaluators relied on their own expertise. In the second phase of the experiment, the evaluation was conducted with the set of ergonomic criteria in the Criteria group and without it in the Control group, but for everyone the evaluation proceeded through the replay of the previous walkthrough. This may have had two consequences: first, in terms of the number of problems uncovered, and second on the nature of the problems that could be detected in the second phase.

The relatively small number of problems uncovered in the first phase in comparison to the total number of problems contained in the system may be due, in part, to time limitation. This is confirmed by the fact that the number of problems uncovered did not differ significantly for participants who explored the modification functionality from those who did not. In other words, whether the participants evaluated all the functionalities of the system or not, the number of problems uncovered did not change significantly. This time limitation may also explain why the proportions of problems found as a function of the number of evaluators in the aggregates are generally less than those reported by Nielsen and Molich (1990).

In the design of this experiment, it was assumed that a second look at the system would entail the discovery of new design flaws whether the participants used the set of ergonomic criteria or not. In other words, it was assumed that spending more time on the evaluation would allow the evaluators of both groups to uncover more problems. It was however hypothesised that the evaluators of the criteria group would uncover significantly more problems during the second phase. The hypothesis was confirmed: in this phase, the evaluators in the criteria group uncovered more problems, and the use of the set of criteria led to only a few errors of classification. The criteria appear to be helpful and usable, at least for usability specialists.

Among the usability problems that appeared to be difficult to uncover are those related to error management, and user control. These results confirm those previously obtained by Nielsen (1992). However, such results can also be explained by the fact that in the second phase of the experiment, the participants could not interact with the system. This could have prevented them from uncovering problems related to error management and user control even when prompted to do so by the criteria.

Nielsen and Molich (1990) indicated deservedly that heuristic evaluation was difficult and that one should not rely on a single evaluation but should rather appeal to three to five evaluators. We believe that given this difficulty, criteria/heuristics should be accompanied by methods to help evaluators be more systematic and exhaustive. By reducing the variability in the evaluation performance, that is by augmenting the similarity between evaluation reports, we would contribute to the efforts dedicated to the design of more economic methods of evaluation. Of course, we are not arguing here that criteria/heuristics evaluation should replace other methods. We believe rather that the set of ergonomic criteria could be used as an organisational framework which would suggest, for each criterion, the evaluation methods most appropriate.

The set of ergonomic criteria has been shown to increase the evaluation performance. However, research work is needed to compare this set with other sets, and to further improve the usefulness and usability of the criteria (e.g., defining a more extensive criteria-based method with detailed procedures) for usability specialists but also for non-specialists. Current work is directed at the latter population of users and at comparing this set to existing standards.

5 References

Apple Computer Inc. (1992). *Macintosh human interface guidelines*. Reading, MA: Addison Wesley.

Bastien, J. M. C., & Scapin, D. L. (1992). A validation of ergonomic criteria for the evaluation of human-computer interfaces. *International Journal of Human-Computer Interaction, 4*, 183-196.

Bastien, J. M. C., & Scapin, D. L. (1993). *Ergonomic criteria for the evaluation of human-computer interfaces* (Technical report No. 156). Rocquencourt, France: Institut National de Recherche en Informatique et en Automatique.

Bodart, F., & Vanderdonckt, J. (in press). *Guide ergonomique de la présentation des applications hautement interactives*. Namur, Belgique: Presses Universitaires de Namur.

Brown, M. H. (1988). Perspectives on algorithm animation. In E. Soloway, D. Frye, & S. B. Sheppard (Eds.), *Proceedings of ACM CHI'88 Conference on Human Factors in Computing Systems* (pp. 33-38). Washington, D.C.: ACM.

de Baar, D. J. M. J., Foley, J. D., & Mullet, K. E. (1992). Coupling application design and user interface design. In P. Bauersfeld, J. Bennett, & G. Lynch (Eds.), *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems* (pp. 259-266). Monterey, California: ACM.

de Souza, F., & Bevan, N. (1990). The use of guidelines in menu interface design: Evaluation of a draft standard. In D. Diaper, D. Gilmore, G. Cockton, & B. Shackel (Eds.), *Proceedings of the IFIP TC 13 Third International Conference on Human-Computer Interaction: INTERACT'90* (pp. 435-440). Cambridge, U.K.: Elsevier Science Publishers.

Desurvire, H., Kondziela, J., & Atwood, M. E. (1992). What is gained and lost when using usability methods other than empirical testing. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems -- Posters and Short Talks* (pp. 125-126). Monterey, California: ACM.

Hammond, N., Hinton, G., Barnard, P., MacLean, A., Long, J., & Whitefield, A. (1985). Evaluating the interface of a document processor: A comparison of expert judgement and user observation. In B. Shackel (Eds.), *Proceedings of the First IFIP Conference on Human-Computer Interaction: INTERACT'84, Vol. 2* (pp. 725-729). London, U.K.: Elsevier Science Publishers.

IBM (1989). *IBM system application architecture. Common User Access: Advanced interface design guide* (Report No. SC26-4582-0). International Business Machines.

Jeffries, R., & Desurvire, H. (1992, October). Usability testing vs. heuristic evaluation: Was there a contest? *SIGCHI Bulletin*, pp. 39-41.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings of ACM CHI'91 Conference on Human Factors in Computing Systems* (pp. 119-124). New Orleans, Louisiana: ACM.

Johnson, G. I., Clegg, C. W., & Ravden, S. J. (1989). Towards a practical method of user interface evaluation. *Applied Ergonomics, 20*, 255-260.

Marshall, C., Nelson, C., & Gardiner, M. M. (1987). Design guidelines. In M. M. Gardiner & B. Christie (Eds.), *Applying cognitive psychology to user-interface design* (pp. 221-278). Chichester: John Wiley & Sons.

Molich, R., & Nielsen, J. (1990, March). Improving a human-computer dialogue. *Communications of the ACM*, pp. 338-348.

Mosier, J. N., & Smith, S. L. (1986). Application of guidelines for designing user interface software. *Behaviour and Information Technology, 5*, 39-46.

Muller, M. J., Dayton, T., & Root, R. (1993). Comparing studies that compare usability assessment methods: An unsuccessful search for stable criteria. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, & T. White (Eds.), *Proceedings of ACM*

INTERCHI'93 Conference on Human Factors in Computing Systems -- Adjunct Proceedings (pp. 185-186). Amsterdam, The Netherlands: ACM.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In P. Bauersfeld, J. Bennett, & G. Lynch (Eds.), *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems* (pp. 373-380). Monterey, California: ACM.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In B. Adelson, S. Dumais, & J. Olson (Eds.), *Proceedings of ACM CHI'94 Conference on Human Factors in Computing Systems, Vol. 1* (pp. 152-158). Boston, Massachusetts: ACM.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In J. Carrasco & J. Whiteside (Eds.), *Empowering people: CHI'90 Conference Proceedings* (pp. 249-256). Seattle, Washington: ACM.

Oppermann, R., Murchner, B., Paetau, M., Pieper, M., Simm, H., & Stellmacher, I. (1989). *Evaluation of dialog systems.* (GMD-Studien No. 169). Gesellschaft Für Mathematik Und Datenverarbeitung MBH.

Pollier, A. (1992). Évaluation d'une Interface par des Ergonomes: Diagnostics et Stratégies [User Interface Evaluation by Human Factors Specialists: Diagnoses and Strategies]. *Le Travail Humain, 55,* 71-96.

Ravden, S. J. (1988). Ergonomic criteria for design of the software interface between human and computer in CIM. *International Journal of Computer Applications in Technology, 1(1-2),* 35-42.

Ravden, S. J., & Johnson, G. I. (1989). *Evaluating usability of human-computer interfaces: A practical method.* Chichester, England: John Wiley & Sons.

Scapin, D. L. (1986). *Guide ergonomique de conception des interfaces homme-machine* (Technical Report No. 77). Rocquencourt, France: Institut National de Recherche en Informatique et en Automatique.

Scapin, D. L. (1990a). Decyphering human factors recommendations. In W. Karwoski & M. Rahimi (Eds.), *Ergonomics of hybrid automated systems II* (pp. 27-34). Amsterdam, The Netherlands: Elsevier Science Publishers.

Scapin, D. L. (1990b). Organizing human factors knowledge for the evaluation and design of interfaces. *International Journal of Human-Computer Interaction, 2,* 203-229.

Scapin, D. L. (1991). *Hypermanip* [HyperCard stack]. Paris: Institut National de Recherche en Informatique et en Automatique.

Shneiderman, B. (1987). *Designing the user interface: Strategies for effective human-computer interaction.* Reading, Massachusetts: Addison-Wesley.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York, NY: McGraw-Hill.

Smith, S. L., & Mosier, J. N. (1986). *Guidelines for designing user interface software* (Report No. ESD-TR-86-278). Mitre Corporation.

Tetzlaff, L., & Schwartz, D. R. (1991). The use of guidelines in interface design. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings of ACM CHI'91 Conference on Human Factors in Computing Systems* (pp. 329-333). New Orleans, Louisiana: ACM.

Thovtrup, H., & Nielsen, J. (1991). Assessing the usability of a user interface standard. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings of ACM CHI'91 Conference on Human Factors in Computing Systems* (pp. 335-341). New Orleans, Louisiana: ACM.

Vanderdonckt, J. M., & Bodart, F. (1993). Encapsulating knowledge for intelligent automatic interaction objects selection. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, & T. White (Eds.), *Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems* (pp. 424-429). Amsterdam, The Netherlands: ACM.

6 Appendices

Table A1. ANOVA table for number of years of experience.

Source	*df*	*Sum of Squares*	*Mean Square*	*F-Value*	*P-Value*
Group	1	9.88	9.88	1	.3316
Subject (Group)	17	168.22	9.90		
Total	18	178.11			

Table A2. Means table for the number of years of experience.

Group	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
Control	10	6.00	3.80	1.20
Criteria	9	4.56	2.19	.73

Table A3. Characteristics of the Participants of Both Groups

		Groups					
		Control (n = 10)		Criteria (n = 9)			
		n	%	n	%	Mo	χ^2
Academic background	Master thesis	1	10.0	0	0.0	***	(4, N = 19) = 2.04 p = .7277
	DESS-Ergo	***5***	***50.0***	***6***	***66.7***		
	DEA	1	10.0	1	11.1		
	DESS & DEA	2	20.0	2	22.2		
	Doctoral thesis	1	10.0	0	0.0		
Percentage of work dedicated to evaluation	10 %	2	20.0	1	11.1	***	(3, N = 19) = .59 p = .8979
	30 %	***5***	***50.0***	***4***	***44.4***		
	50 %	2	20.0	3	33.3		
	70 %	1	10.0	1	11.1		
	100 %	0	00.0	0	00.0		
Evaluation of specifications	never	0	0.0	0	0.0		(1, N = 18) = 2.81 p = .0935
	sometimes	5	50.0	***7***	***87.5***		
	often	5	50.0	1	12.5		
Evaluation of screen dumps	never	3	30.0	0	0.0	***	(2, N = 18) = 4.65 p = .0976
	sometimes	***4***	***40.0***	***7***	***87.5***		
	often	3	30.0	1	12.5		
Evaluation of prototypes	never	1	10.0	0	0.0	***	(2, N = 17) = 3.66 p = .1603
	sometimes	***6***	***60.0***	***7***	***100.0***		
	often	3	30.0	0	0.0		
Evaluation of mockups	never	0	0.0	0	0.0		(1, N = 19) = .42 p = .5146
	sometimes	3	30.0	4	44.4		
	often	***7***	***70.0***	***5***	***55.6***	***	
Evaluation of operational system	never	2	20.0	3	37.5		(2, N = 18) = .69 p = .7095
	sometimes	***5***	***50.0***	3	37.5		
	often	3	30.0	2	25.0		
Mac users	yes	***10***	***100.0***	***9***	***100.0***	***	
	no	0	0.0	0	0.0		
HyperCard experience	no	3	30.0	3	33.3		(3, N = 19) = 2.29 p = .515
	user	***4***	***40.0***	1	11.1		
	programmer	1	10.0	2	22.2		
	user & progra	2	20.0	3	33.3		
Database experience	no	2	20.0	3	33.3		(5, N = 19) = 3.36 p = .6452
	user	2	20.0	0	0.0		
	evaluator	2	20.0	3	33.3		
	user & eval.	2	20.0	2	22.2		
	eval. & design	1	10.0	0	0.0		
Evaluation based on expertise	not at all	0	0.0	0	0.0		(1, N = 18) = 1.94 p = .1632
	a little	1	10.0	3	37.5		
	a lot	***9***	***90.0***	***5***	***62.5***	***	
User testing	not at all	0	0.0	0	0.0		(1, N = 19) = .46 p = .4977
	a little	4	40.0	***5***	***55.6***		
	a lot	***6***	***60.0***	4	44.4		
Use of guides / recommendations	not at all	2	20.0	1	11.1	***	(2, N = 19) = .69 p = .7072
	a little	***6***	***60.0***	***7***	***77.8***		
	a lot	2	20.0	1	11.1		
Use of checklists	not at all	3	30.0	1	12.5	***	(2, N = 18) = .79 p = .6745
	a little	***6***	***60.0***	***6***	***75.0***		
	a lot	1	10.0	1	12.5		

Note. Numbers in boldface italics indicate the mode. The three asterisks indicate questionnaire items for which the mode was the same for both groups. Even though we must be cautious about the Chi squares computed due to the small cells frequencies in this table (Siegel, 1956), they all give indication that the two groups did not differ on any of the questionnaire items

Table A4. Multivariate ANOVA for repeated measures for the number of problems uncovered: Group and Phase effects.

Source	*df*	*Sum of Squares*	*Mean Square*	*F-Value*	*P-Value*	*G-G*	*H-F*
Groupe	1	1408.78	1408.78	5.59	.0302		
Subject(Group)	17	4280.69	251.81				
Phase	1	14561.77	14561.77	333.85	.0001	.0001	0001
*Phase * Groupe*	1	451.77	451.77	10.36	.0050	.0050	.0050
*Phase * Subject(Group)*	17	741.49	43.62				

Table A5. Means table for the number of problems uncovered: Group effect.

Group	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
Control	20	77.75	20.70	4.63
Criteria	18	89.94	26.20	6.17

Table A6. Means table for the number of problems uncovered: Phase effect.

	Count	*Mean*	*Std. Dev.*	*Std. Error*
Phase 1	19	64.11	10.87	2.49
Phase 2	19	102.95	16.26	3.73

Table A7. Means table for the number of problems uncovered: Phase x Group effects.

	Count	*Mean*	*Std. Dev.*	*Std. Error*
Phase 1, Control	10	61.60	10.48	3.31
Phase 1, Criteria	9	66.89	11.21	3.74
Phase 2, Control	10	93.90	14.66	4.64
Phase 2, Criteria	9	113.00	11.69	3.90

Table A8. Univariate ANOVA table for the number of problems uncovered: Group and Phase effects.

Source	*df*	*Sum of Squares*	*Mean Square*	*F-Value*	*P-Value*
Groupe	1	1408.78	1408.78	5.59	.0302
Subject(Group)	17	4280.69	251.81		
Phase	1	14561.77	14561.77	333.85	.0001
*Phase * Groupe*	1	451.77	451.77	10.36	.0050
*Phase * Subject(Group)*	17	741.49	43.62		

Table A9. Interaction contrast table for groups in phase 1: Phase * Group effect. Dependent: Number of problems uncovered.

	Cell Weight
Phase 1, Control	1.00
Phase 1, Criteria	-1.00
df	1
Sum of Squares	132.50
Mean Square	132.50
F-Value	3.04
P-Value	.0994

Table A10. Interaction contrast table for groups in phase 2: Phase * Group effect. Dependent: Number of problems uncovered.

	Cell Weight
Phase 2, Control	1.00
Phase 2, Criteria	-1.00
df	1
Sum of Squares	1728.05
Mean Square	1728.05
F-Value	39.62
P-Value	.0001

Table A11. Interaction contrast table for phases in the Control group: Phase * Group effect. Dependent: Number of problems uncovered.

	Cell Weight
Phase 1, Control	1.00
Phase 2, Control	-1.00
df	1
Sum of Squares	5216.45
Mean Square	5216.45
F-Value	119.60
P-Value	.0001

Table A12. Interaction contrast table for phases in the Criteria group: Phase * Group effect. Dependent: Number of problems uncovered.

	Cell Weight
Phase 1, Criteria	1.00
Phase 2, Criteria	-1.00
df	1
Sum of Squares	9568.06
Mean Square	9568.06
F-Value	219.36
P-Value	.0001

Table A13. Multivariate ANOVA table for repeated measures for the number of problems uncovered: Completion of the first phase, Group, and Phase effects.

Source	*df*	*Sum of Squares*	*Mean Square*	*F-Value*	*P-Value*	*G-G*	*H-F*
Completion of phase 1	1	644.60	644.60	2.73	.1190		
Group	1	1063.48	1063.48	4.51	.0507		
*Completion * Group*	1	49.78	49.78	.21	.6525		
Subject(Group)	15	3537.33	235.82				
Phase	1	11377.78	11377.78	278.75	.0001	.0001	.0001
*Phase * Completion*	1	47.46	47.46	1.16	.2979	.2979	.2979
*Phase * Group*	1	455.11	455.11	11.15	.0045	.0045	.0045
*Phase * Completion * Group*	1	67.60	67.60	1.66	.2176	.2176	.2176
*Phase * Subject(Group)*	15	612.25	40.82				

Table A14. Means table for the number of problems uncovered: Completion effect.

Completion	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
yes	28	85.50	23.10	4.37
no	10	78.00	26.65	8.43

Table A15. Means table for the number of problems uncovered: Completion x Group effects.

Completion, Group	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
yes, Control	16	79.12	21.00	5.25
yes, Criteria	12	94.00	23.88	6.89
no, Control	4	72.25	21.38	10.69
no, Criteria	6	81.83	30.99	12.65

Table A16. Means table for the number of problems uncovered: Phase x Completion effects.

Phase, Completion	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
Phase 1, yes	14	67.14	10.17	2.72
Phase 1, no	5	55.60	8.53	3.82
Phase 2, yes	14	103.86	16.71	4.47
Phase 2, no	5	100.40	16.46	7.36

Table A17. Means table for the number of problems uncovered: Phase x Completion x Group effects.

Phase, Completion, Group	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
Phase 1, yes, Control	8	62.88	11.23	3.97
Phase 1, yes, Criteria	6	72.83	4.92	2.01
Phase 1, no, Control	2	56.50	6.36	4.50
Phase 1, no, Criteria	3	55.00	11.14	6.43
Phase 2, yes, Control	8	95.38	14.69	5.19
Phase 2, yes, Criteria	6	115.17	12.46	5.09
Phase 2, no, Control	2	88.00	18.38	13.00
Phase 2, no, Criteria	3	108.67	10.79	6.23

Table A18. Multivariate ANOVA table for repeated measures for the number of problems uncovered: Exploration of the modification functionality, Group, and Phase effects.

Source	*df*	*Sum of Squares*	*Mean Square*	*F-Value*	*P-Value*	*G-G*	*H-F*
Explor. Modification	1	796.45	796.45	3.45	0829		
Group	1	1856.46	1856.46	8.04	.0125		
*Explor. Modification * Group*	1	58.08	58.08	.25	.6232		
Subject(Group)	15	3461.40	230.76				
Phase	1	7513.89	7513.89	163.70	.0001	.0001	.0001
*Phase * Explor. Modification*	1	2.57	2.57	.06	.8161	.8161	.8161
*Phase * Group*	1	285.06	285.06	6.21	.0249	.0249	.0249
*Phase * Exp. Modif. * Group*	1	33.10	33.10	.72	.4091	.4091	.4091
*Phase * Subject(Group)*	15	688.50	45.90				

Table A19. Means table for the number of problems uncovered: Exploration of the modification functionality effect.

Explo. Modif.	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
oui	12	82.17	27.06	7.81
non	26	84.15	22.91	4.49

Table A20. Means table for the number of problems uncovered: Explor. Modification x Group effects.

Exp. Modif. Group	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
yes, Control	2	63.50	20.51	14.50
yes, Criteria	10	85.90	27.49	8.69
no, Control	18	79.33	20.68	4.87
no, Criteria	8	95.00	25.35	8.96

Table A21. Means table of the number of problems uncovered: Phase x Explor. Modification effects.

Phase, Modif.	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
Phase 1, yes	6	59.33	11.55	4.72
Phase 1, no	13	66.31	10.24	2.84
Phase 2, yes	6	105.00	15.05	6.14
Phase 2, no	13	102.00	17.29	4.80

Table A22. Means table for the number of problems uncovered: Phase x Exploration of the Modification x Group effects.

Phase, Exp. Modifi. Group	*Count*	*Mean*	*Std. Dev.*	*Std. Error*
Phase 1, yes, Control	1	49.00	•	•
Phase 1, yes, Criteria	5	61.40	11.61	5.19
Phase 1, no, Control	9	63.00	10.07	3.36
Phase 1, no, Criteria	4	73.75	6.50	3.25
Phase 2, yes, Control	1	78.00	•	•
Phase 2, yes, Criteria	5	110.40	8.02	3.59
Phase 2, no, Control	9	95.67	14.38	4.79
Phase 2, no, Criteria	4	116.25	15.92	7.96

Table of contents

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine - Technopôle de Nancy-Brabois - Campus scientifique
615 rue du Jardin Botanique - B P 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes - IRISA, Campus universitaire de Beaulieu 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes 46 avenue Félix Viallet - 38031 Grenoble Cedex 1 (France)
Unité de recherche INRIA Sophia Antipolis - 2004 route des Lucioles - B P 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt - B P 105 - 78153 Le Chesnay Cedex (France)

ISSN 0249 - 6399